IP Layer Metrics for 5G Edge Computing Service

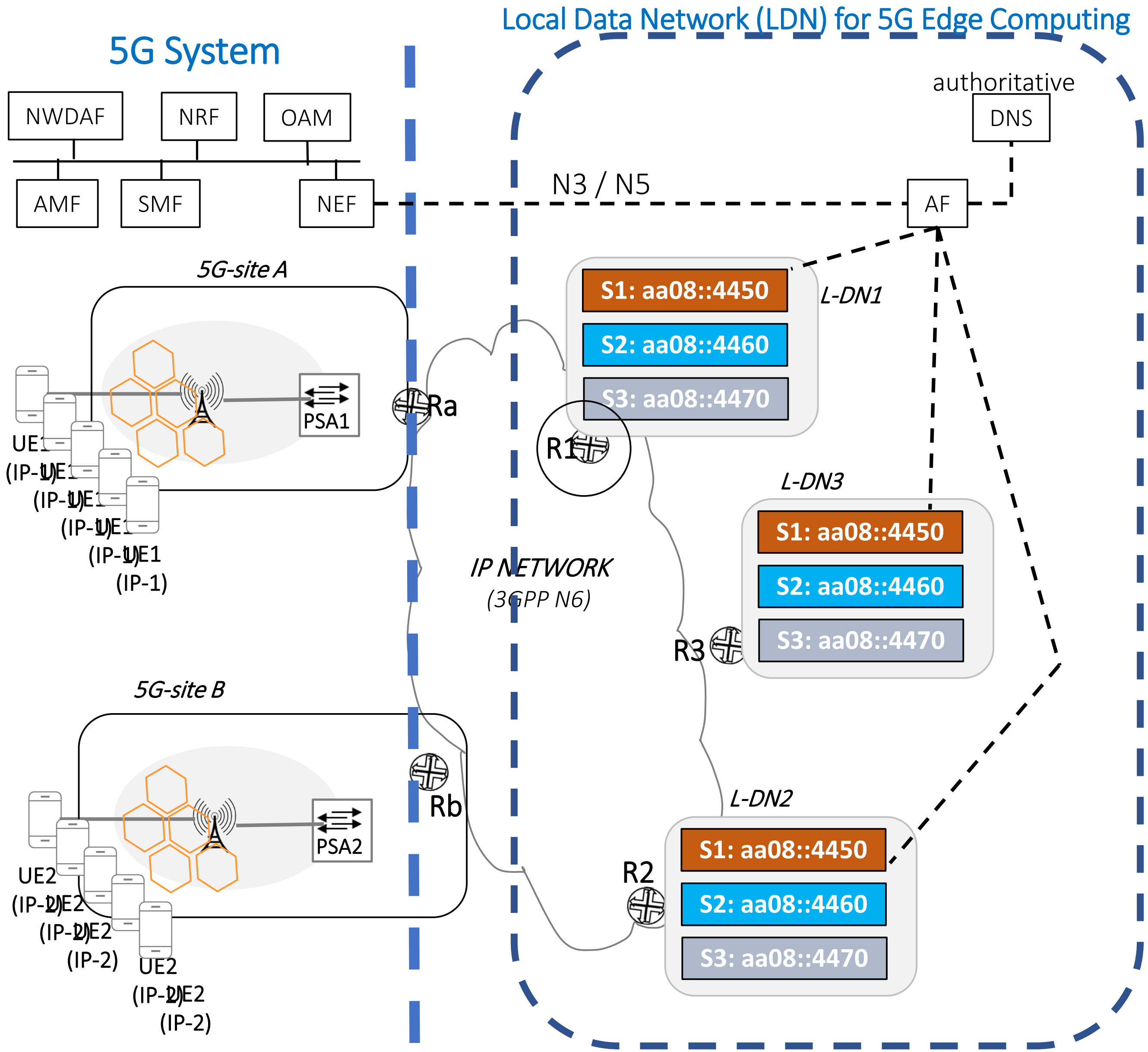
draft-dunbar-ippm-5g-edge-compute-ip-layer-metrics-01

Linda Dunbar

HaoYu Song

John Kaippallimalil

5G Edge Computing (3GPP TR23.748)

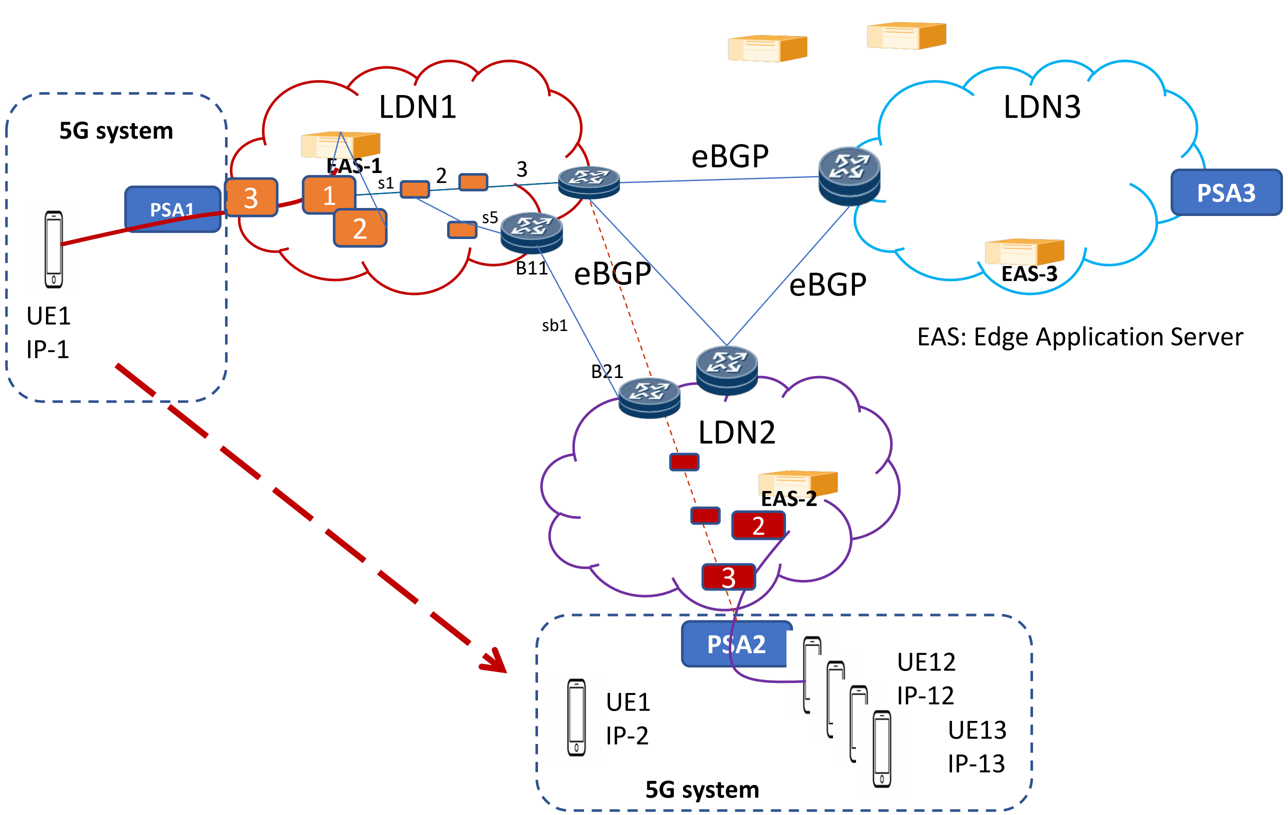


One Application has multiple Application Servers located in Edge Computing DCs

From IP Network Perspective…

ANYCAST: IP Layer Application ID -> multiple App servers
Benefit of using ANYCAST:

- ✓ **dynamically load balance across locations based on network conditions.**
- ✓ **leverages the proximity information present in the network (routing) layer and**
- ✓ **eliminates the single point of failure and bottleneck at the DNS resolvers and application layer load balancers.**
- ✓ **removes the dependency on UEs using their cached destination IP addresses for extended period**



EAS: Edge Application Server

Problem 1: Selecting 5G Edge Application Location

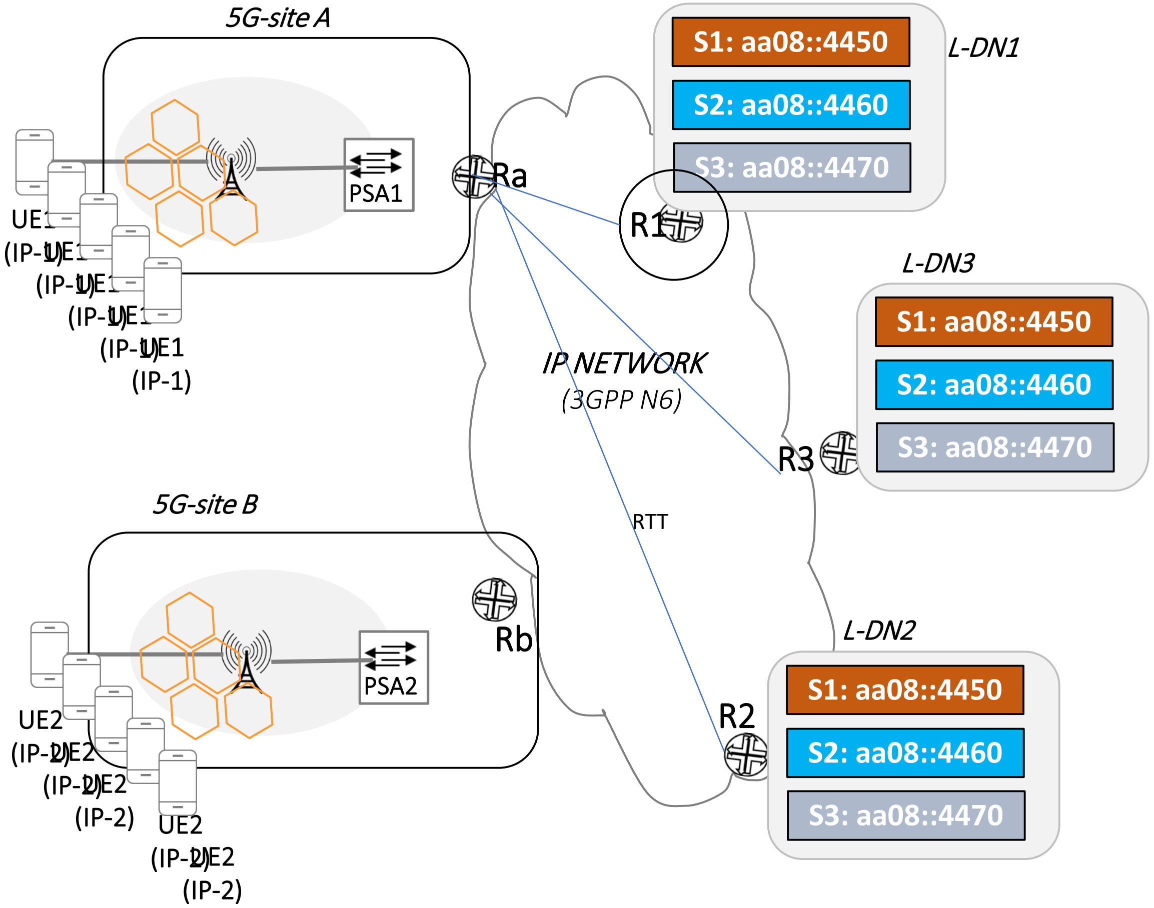
- ➢ Many mini data centers can be close in proximity, making it difficult to differentiate in Routing Hops for App servers hosted in them,
- ➢ Some data centers can have higher capacity than others,
- ➢ Some sites may be more preferred when a UE anchored to a new 5G Site

Problem 2: UE mobility creates unbalanced anycast distribution

RTT to an ANYCAST Address in 5G EC

- RTT to "app.net" ANYCAST S1:
- List of {
  - R1: RTT value
  - R2: RTT value
  - R3: RTT value

  }



Algorithm in Selecting the Optimal Target Location

To compare the cost to reach the Application Server between the Site-i or Site-j:

```
                 Load-i * CP-j                Pref-j * Delay-i
Cost-i=min(w *(----------------) + (1-w) *(------------------))
                 Load-j * CP-i                Pref-i * Delay-j
```

- Load-i: Load Index at Site-I = w1*ToPackets+w2*FromPackes+w3*ToBytes+w4*FromBytes

  0<= wi <=1 and w1+ w2+ w3+ w4 = 1.

- CP-i (Capacity-i) (higher value means higher capacity): capacity index at the site i.
- Delay-i: Network latency measurement (RTT) to the A-ER that has the Application Server attached at the site-i.
- Pref-i (Preference Index: higher value means higher preference): Network Preference index for the site-I.
- w: Weight for load and site information,
  - 0<= w <=1: If smaller than 0.5, Network latency and the site Preference have more influence; otherwise, Server load and its capacity have more influence.

Next step:

- To standardize the IP Layer Metrics for Application Servers
  - To make network more aware of application server running status and environment
  - To achieve more optimized delivery of service
- Need your feedback